Machine Learning and Data Mining for Improved Intelligent Data Understanding of High Dimensional Earth Science Data

Prof. Carla Brodley
School of Electrical and Computer Engineering
Purdue University

Prof. Mark Friedl
Department of Geography and Center for Remote Sensing
Boston University

Challenge 1: Multiple Class and Highly Skewed Class Distribution

- For example: the smallest class in MODIS data set is 0.8% of the labeled dataset
- Minority classes are difficult to handle for classifiers
- A large number of classes is difficult to handle for state of the art data mining methods

Reducing Multiclass to Binary via Class Elimination

Hribar, Fern, and Brodley, submitted to the *Twenty-First International Conference on Machine Learning*

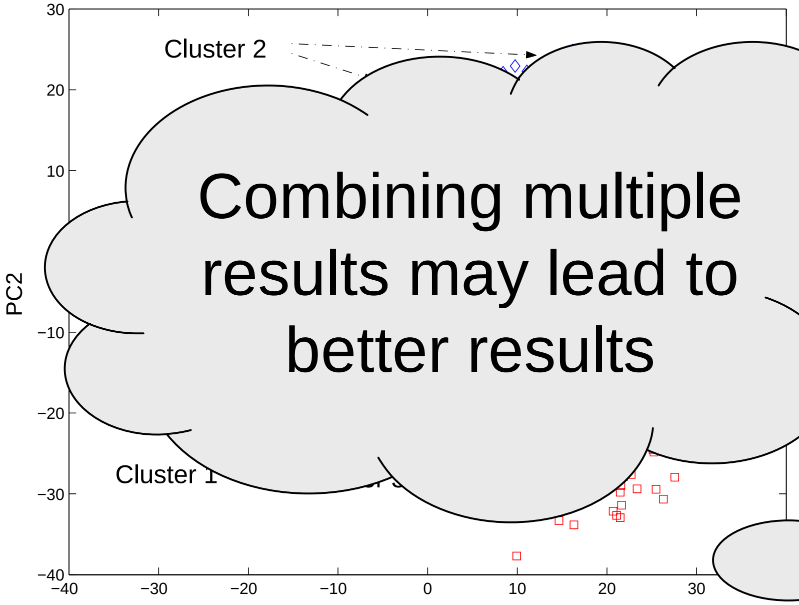
- Step 1: Reduce multiclass to binary via class elimination
- Step 2: Apply a binary classifier trained on just those two classes
- Implementation: decision trees for elimination, SVM for binary

- Results:
  - – 5-8% improvement in overall accuracy
  - – Large improvements in minority class accuracy

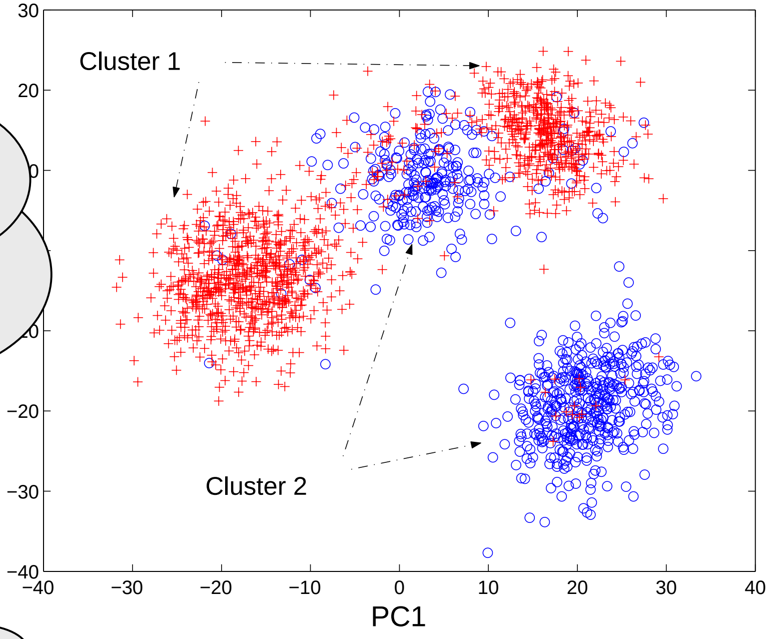
Challenge 2: Forming Clusters of High-Dimensional Data

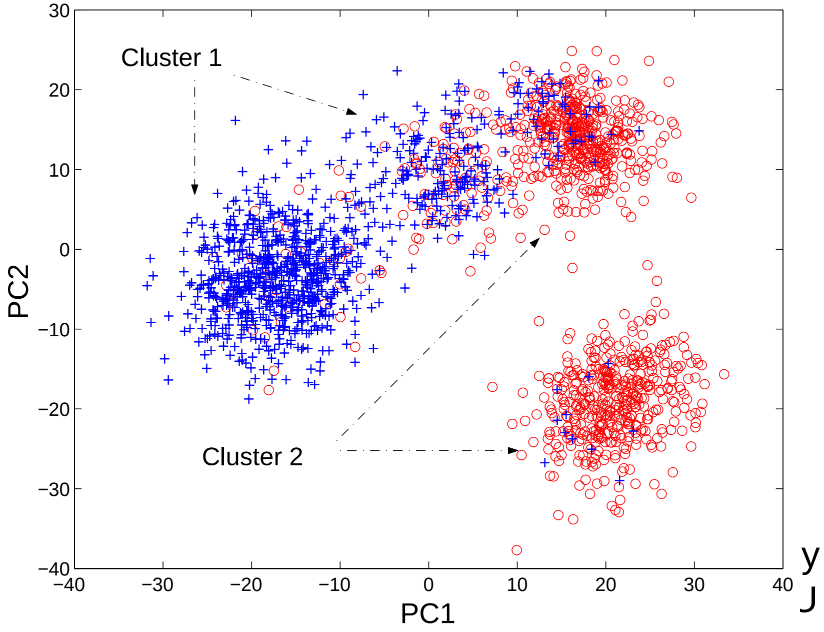
- Difficult for current algorithms
- Projection techniques make assumptions about the data

Examples of RP Clustering Results



Combining multiple results may lead to better results



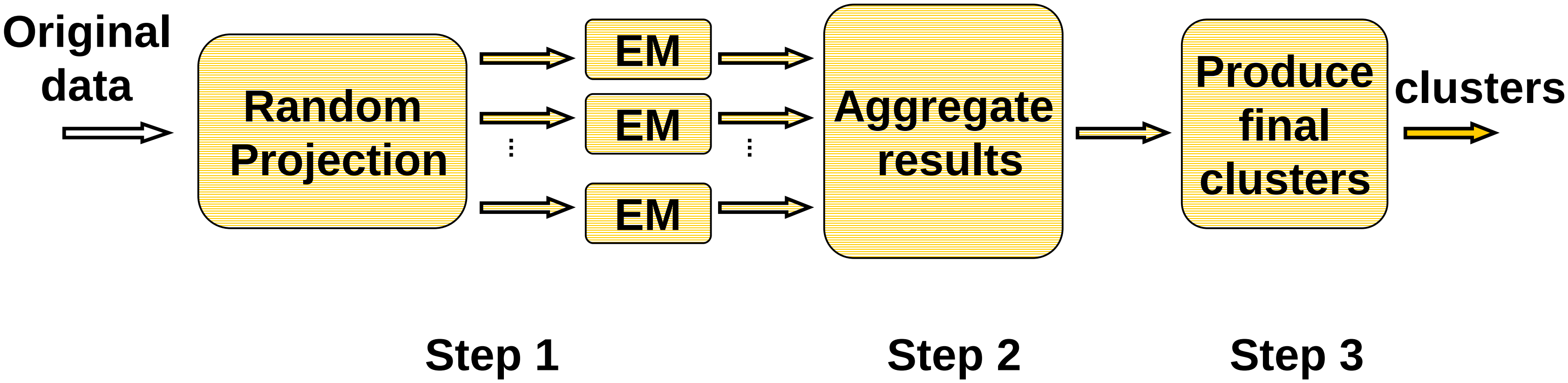


Different runs reveal different structure

y and Friedl
J Workshop 04

Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach,

Fern and Brodley, *Proceedings of the Twentieth International Conference on Machine Learning*, 2003

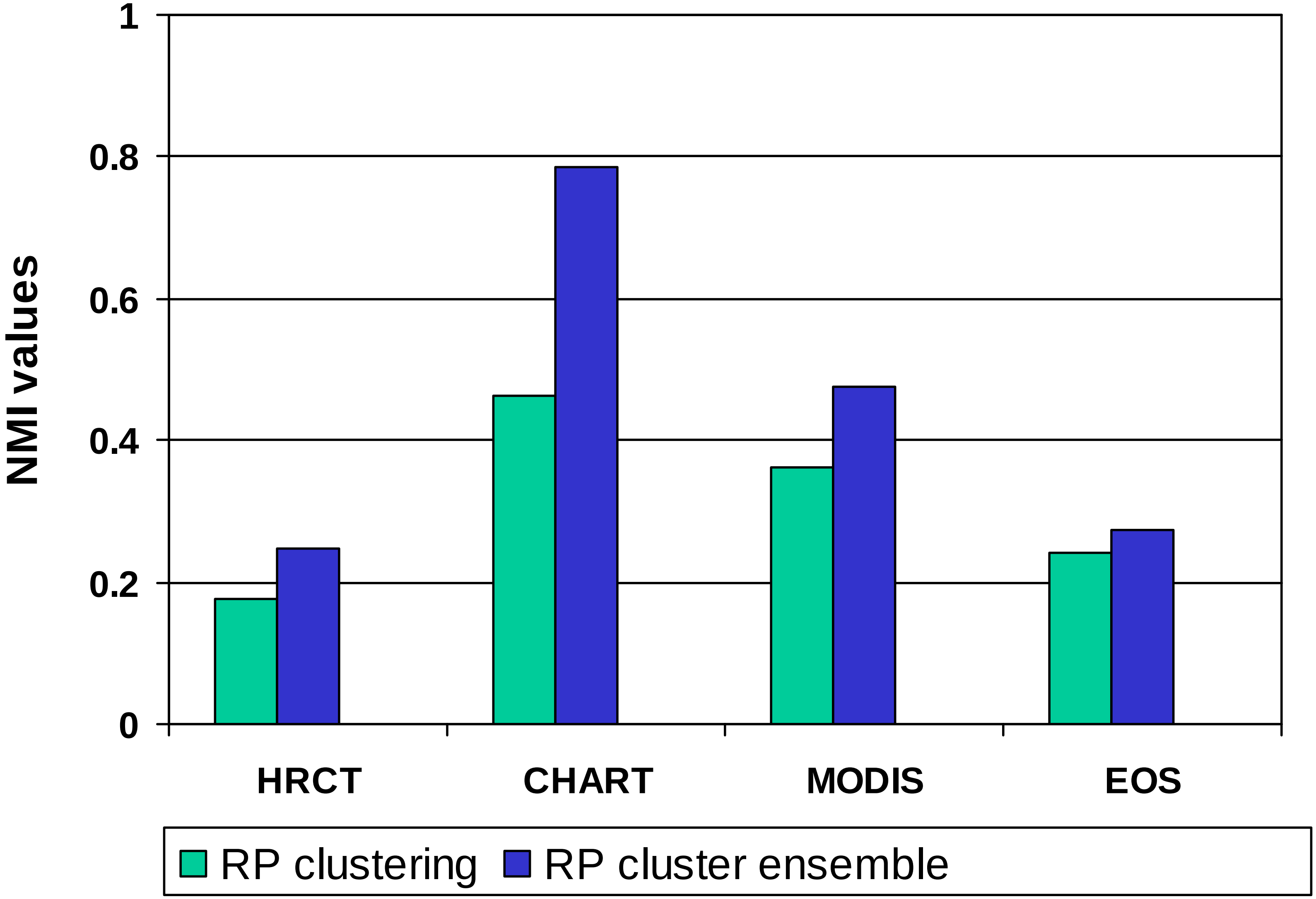


Step 1. Generate multiple RP clustering results

Step 2. Aggregate the results

Step 3. Produce final clusters

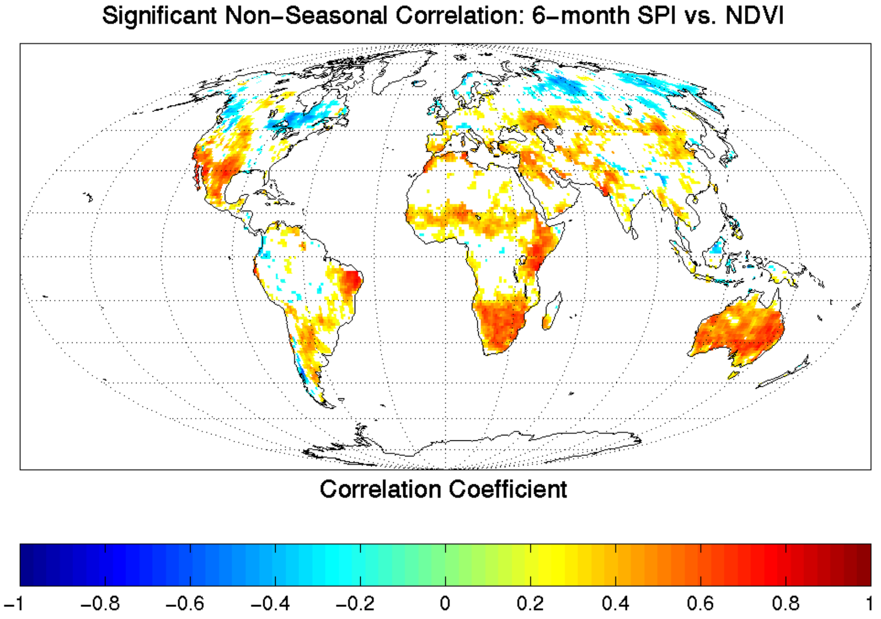
Experimental Results



Application of Data Mining to Earth Science Data Sets

Two main goals

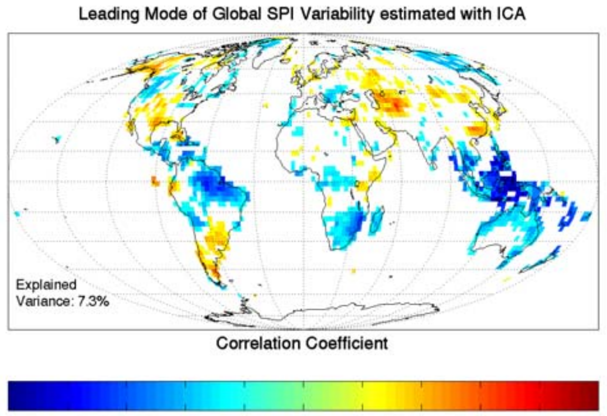
1. Implementation and evaluation of tools for *intelligent understanding of time series image data*
2. Discovery of *climate-ecosysten interactions* in support of NASA's Earth Science Enterprise

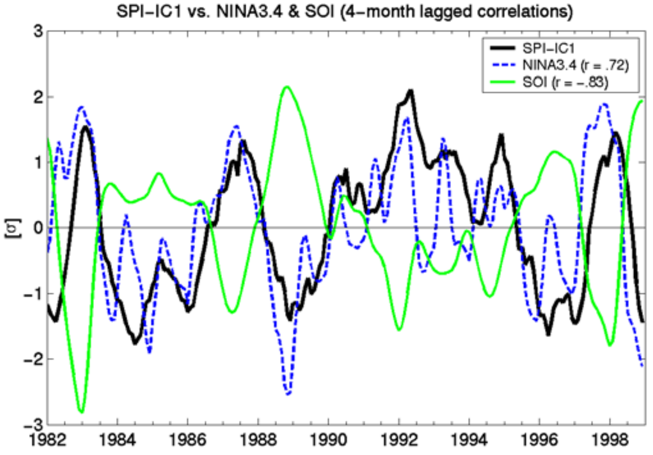


Application of Data Mining to Earth Science Data Sets

Three main activities

1. Non-linear decomposition of time series NDVI and SST images
2. Analysis of non-seasonal co-variability in precipitation and vegetation dynamics
3. Analysis of coupled non-seasona SST, precipitation, and NDVI anomalies

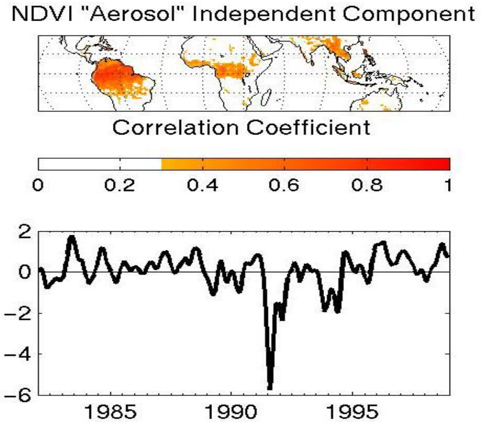


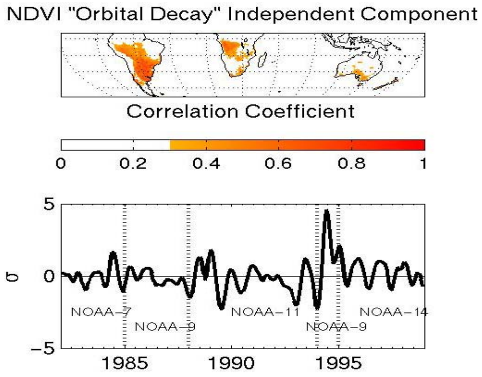


Non-Linear Decomposition of Time Series NDVI and SST Images

Lotsch et al., 2003. *IEEE Trans. on Geosci. and Rem. Sens.*, 41(12): 2938-2942

- Analysis of spatio-temporal variance
  - Conventionally use linear methods
- For this work: ICA
  - NDVI and SST image time series
- Results reveal additional information not identified by linear methods
  - Artifacts in NDVI data sets
  - Decoupled “modes” of variation in SST related to ENSO and PDO


NDVI "Aerosol" Independent Component

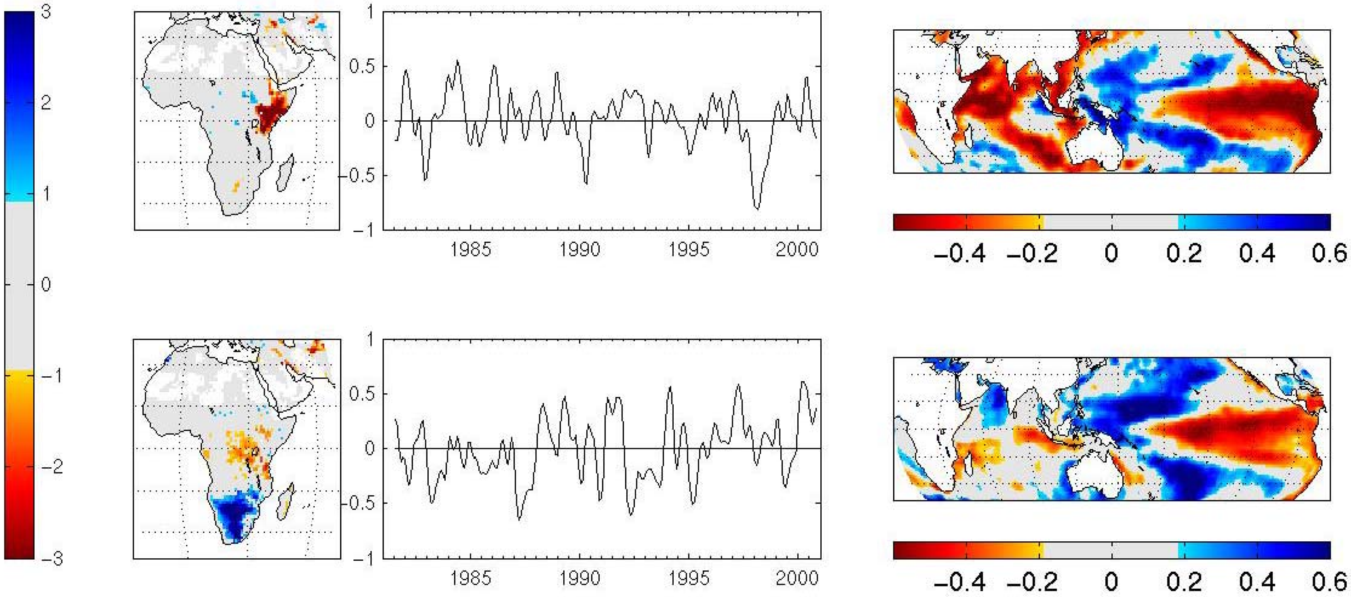

NDVI "Orbital Decay" Independent Component

Covariability in Non-Seasonal Precipitation & Ecosystem Dynamics

Lotsch et al. 2003. *Geophysical Research Letters*, 30(14), 1774
doi: 10.1029/2003GL01756

- Remove seasonal variation
  - Dominates variance
- Investigate signature of climate forcing at interannual time scales

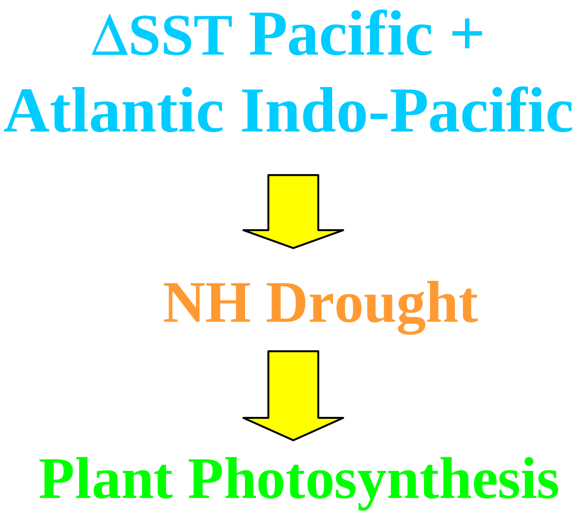
- Canonical Correlation Analysis
  - NDVI and SPI
- Analysis reveals regional patterns of NDVI-SPI variation associated with specific forcing mechanisms



Analysis of Coupled SST, Precipitation, & NDVI Anomalies

Lotsch et al, 2004, in prep for *Bull. Am. Met Soc*.

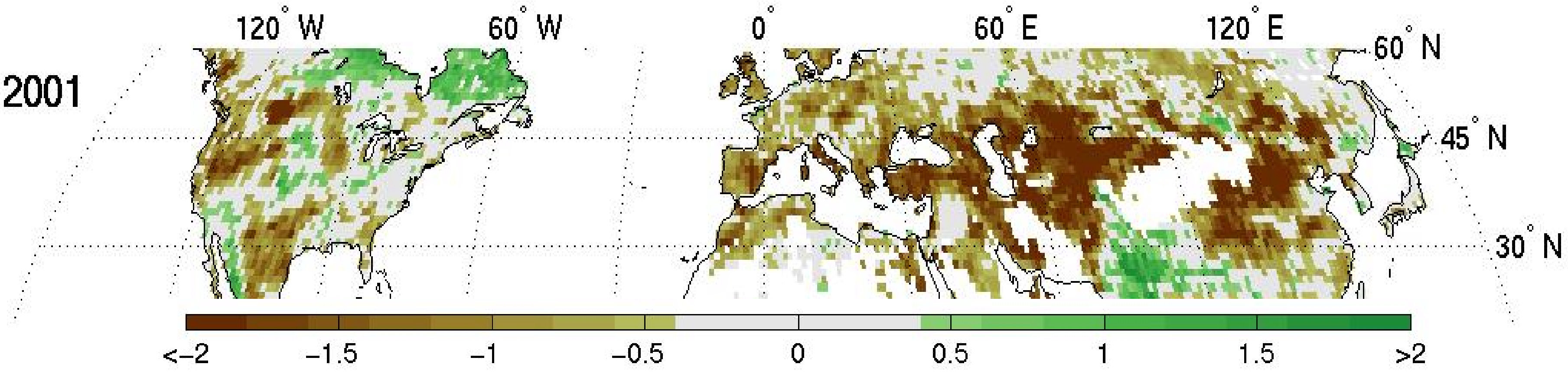
- Dramatic reduction in plant growth linked to ***synchronous patterns*** of SST fluctuations and geographically extensive precipitation anomalies in Northern Hemisphere mid-latitudes during 1998-2002



ΔSST Pacific + Atlantic Indo-Pacific

NH Drought

Plant Photosynthesis



Summary and Expected Impact on NASA
(Supervised and Unsupervised Research)

- Adaptation of innovative new data mining methods to space-time data sets for climate and geophysical data.
  - Provides Earth Science Enterprise activities with a new method for examining and understanding large volume, high dimensional remote sensing and geophysical datasets
- Development of new data mining methods for handling multiple classes, minority classes and missing data
  - provides better methods for generating accurate, complete, global land cover maps.

Research Plans

- Apply nonlinear CCA to Earth science data sets, in preparation for KDD 04
- Apply class elimination ideas to land cover classification, in preparation for the *Journal of Machine Learning Research*
- Complete journal article on ensemble methods for submission to *Journal of Machine Learning Research*

Personnel

- Carla Brodley, Co-PI
- Mark Friedl, Co-PI
- Xiaoli Z. Fern, Ph.D. student, Purdue Univ.
- Nate Hribar, M.S. student, Purdue Univ.
- Alex Lotsch, Ph.D. student, Boston Univ.
- Su-Yin Tan M.A. student, Boston Univ.

References

- Anderson, B.T., Lotsch,A. and M.A. Friedl,2002. Using independent component analysis for non-linear decorrelation of SST modes. *AGU*, 83(47) Fall Meeting Suppl., Abstract NG72A-0920. Dec 6-10, 2002, San Francisco, CA.

- Dy, J. and Brodley, C. E., ``Feature selection for unsupervised learning,'' accepted to appear in *Journal of Machine Learning Research.*

- Fern, X. Z. and Brodley, C. E., ``Boosting lazy decision trees,'' *Proceedings of the Twentieth International Conference on Machine Learning*, August 2003 Washington D.C.

- Fern, X. Z. and Brodley, C. E., ``Random projection for high dimensional data clustering: A cluster ensemble approach,'' *Proceedings of the Twentieth International Conference on Machine Learning*, August 2003 Washington D.C.

- Fern, X. Z. and Brodley, C. E., ``Solving cluster ensemble problems by bipartite graph partitioning,'' submitted to the *Twenty-First International Conference on Machine Learning.*

- Hribar, N., Fern, X. Z. and Brodley, C. E., ``Reducing multiclass to binary via class elimination,'' submitted to the *Twenty-First International Conference on Machine Learning.*

- Lotsch, A, M.A. Friedl, and J. Pinzon, 2003. Spatio-Temporal Deconvolution of NDVI Image sequences using independent component analysis, *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 41. No. 12, pp. 2938-2942*.*

- Lotsch, A., Friedl, M.A., Anderson, B.T. and C.J. Tucker 2003. Coupled vegetation-precipitation variability observed from satellite and climate records, *Geophysical Research Letters,* 30(14), 1774, doi: 10.1029/2003GL017506

- Lotsch, A., Friedl, M.A., Anderson, B.T. and C.J. Tucker 2004. Response of terrestrial ecosystems to recent northern hemisphere drought, in preparation for submission to *Bulletin of the American Meteorological Society*, Feb 2004.

- Lotsch, A, M.A. Friedl, B.T. Anderson, and C.J. Tucker, 2003. Linking ocean-atmosphere dynamics to precipitation-vegetation covariability, *Eos. Trans. AGU*, 84(46) Fall Meeting Suppl., Abstract B52E-05. Dec 8-12, 2003, San Fran., CA.

- Lotsch, A. and M.A. Friedl 2002. Using linear and non-linear methods to study precipitation-vegetation dynamics at global scales, *Eos. Trans. AGU*, 83(47) Fall Meeting Suppl., AbstractB21B-0730. Dec 6-10, 2002, San Francisco, CA.

Machine Learning and Data Mining for Improved Intelligent Data Understanding of High Dimensional Earth Science Data

Carla Brodley/Purdue and Mark Friedl/Boston University

Goal: Knowledge discovery for large multivariate Earth science datasets.

Objectives: Develop computationally efficient machine learning algorithms for intelligent data understanding for large multivariate Earth science datasets.

Key Innovations:

- Clustering ensembles for unsupervised learning of high-dimensional data
- Solving the unbalanced multiclass learning problem
- Application of non-linear decomposition methods to time series image data
- Discovery of joint climate ecosystem co-variability via data mining

NASA Relevance:

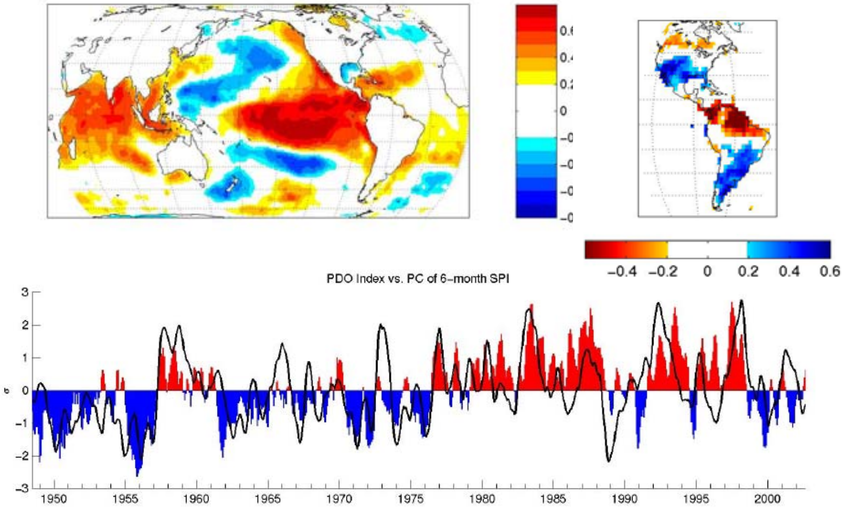
- NASA Earth science enterprise requires efficient methods for knowledge discovery in global multivariate time series data.
- Extension to astrophysics and homeland security.

Accomplishments to date:

- 3 refereed journal papers, 2 refereed conference papers, 3 conference abstracts, 3 papers in prep.

Sea surface temperature anomalies

Ecosystem response



Covariance in ocean/precipitation dynamics

Schedule:

- FY01: Data set compilation; application of ICA to NDVI and sea surface temp. time series; developed lazy decision tree and class elimination algorithms.
- FY02: Analysis of climate ecosystem co-variability; developed cluster ensemble framework
- FY03: Analysis of ocean-atmosphere ecosystem teleconnections; refinement of cluster ensemble, class elimination algorithms